UNIVERSITA' DEGLI STUDI DI TRENTO - DIPARTIMENTO DI ECONOMIA

A FRAMEWORK FOR CUT-OFF SAMPLING IN BUSINESS SURVEY DESIGN

Marco Bee
Roberto Benedetti
Giuseppe Espa

Discussion Paper No. 9, 2007

The Discussion Paper series provides a means for circulating preliminary research results by staff of or visitors to the Department. Its purpose is to stimulate discussion prior to the publication of papers.

Requests for copies of Discussion Papers and address changes should be sent to:

Dott. Stefano Comino
Dipartimento di Economia
Università degli Studi
Via Inama 5
38100 TRENTO ITALIA

A Framework for Cut-off Sampling in Business Survey Design

Marco Bee

Department of Economics, University of Trento, 38100 Trento, Italy

Roberto Benedetti

Department of Business, Statistical, Technological and Environmental Sciences (DASTA), University "G. d'Annunzio" of Chieti-Pescara

Giuseppe Espa

Department of Economics, University of Trento, 38100 Trento, Italy

Summary. In sampling theory the large concentration of the population with respect to most surveyed variables constitutes a problem which is difficult to tackle by means of classical tools. One possible solution is given by cut-off sampling, which explicitly prescribes to discard part of the population; in particular, if the population is composed by firms or establishments, the method results in the exclusion of the "smallest" firms. Whereas this sampling scheme is common among practitioners, its theoretical foundations tend to be considered weak, because the inclusion probability of some units is equal to zero. In this paper we propose a framework to justify cut-off sampling and to determine the census and cut-off thresholds. We use an estimation model which assumes as known the weight of the discarded units with respect to each variable; we compute the variance of the estimator and its bias, which is caused by violations of the aforementioned hypothesis. We develop an algorithm which minimizes the MSE as a function of multivariate auxiliary information at the population level. Considering the combinatorial optimization nature of the model, we resort to the theory of stochastic relaxation: in particular, we use the simulated annealing algorithm.

Keywords: Cut-off sampling, skewed populations, model-based estimation, optimal stratification, simulated annealing

1. Introduction

Cut-off sampling is a procedure commonly used by national statistical institutes to select samples, but it is not easy to give a unique, clear-cut definition of the methodology. Roughly speaking, the population is partitioned in two or three strata such that the units in each stratum are treated differently; in particular, part of the target population is usually excluded a priori from sample selection.

The basic formulation (Hansen *et al.* 1953, pagg. 486-490, Särndal *et al.* 1992, pagg. 531-533), frequently employed in the field of price collection, is characterized by a threshold such that the units above this threshold are included in the sample with probability one and the units below the threshold are discarded, namely their probability of being included in the sample is zero. In this case, as noted by Haan *et al.* (1999), the sampling variance is zero by definition.

An alternative interpretation is proposed by Hidiroglou (1986), who considers a stratum where, as before, all the observations are included in the sample, and a second stratum where the units are not discarded but sampled.

Finally, the most general approach (the one adopted in this paper) considers three strata whose units are respectively enumerated completely, sampled and discarded.

As pointed out by Sigman and Monsour (1995), this type of stratification is particularly appropriate in business surveys, because businesses tend to have skewed distributions with many small units and very few large units. Thus, size has a considerable impact on the precision of survey estimates, and failure to notice that such populations should be stratified in the aforementioned manner may cause an underestimation of the population characteristics. When the distribution of the selection variable is concentrated in few large establishments, this methodology provides the investigator with a sample whose size is rather small but whose degree of coverage is high.

The problem treated in this paper is a generalization of standard cut-off sampling. Therefore, as usual in business surveys, we assume that the population of interest is positively skewed, because of the presence of few "large" units and many "small" units. If the investigator is interested in estimating the total of the population, a considerable percentage of the observations gives a negligible contribution to the estimate of the total. On the other hand, the inclusion in the sample of the largest observations is essentially mandatory.

In such situations, practitioners often use partitions of the population in three sets: a take-all stratum whose units are surveyed entirely (U_C), a take-some stratum from which a simple random sampling is drawn (U_S) and a take-nothing stratum whose units are discarded (U_E). In other words, survey practitioners decide a priori to exclude from the analysis part of the population (for example, firms with less than five employees); however, this choice is often motivated by the desire to match administrative rules (in this case, the partition of firms in small, medium and large). This strategy is employed so commonly in business surveys that its use is "implicit" and "uncritical"; the inferential consequences of the restrictions caused to the archive by this procedure are mostly ignored.

The problem of finding the optimal take-all threshold, i.e. the partition of the population in strata U_C and U_S, is relatively straightforward both from the technical and from the methodological point of view (Hidiroglou 1986). On the other hand, finding a criterion

which assigns each unit to exactly one of the three strata tends to be considered as a non-viable alternative, mainly because some inclusion probabilities are set equal to zero. It follows that cut-off sampling is, in some sense, in an intermediate position between probabilistic and non-probabilistic sampling schemes, a feature which is not appreciated by experts in this field. As a result, in the literature there are very few papers concerning its methodological foundations.

Nonetheless, in applications it is frequently used; it is the case, for example, of the monthly survey of manufacturing performed by Statistics Canada (see, for example, Statistics Canada 2001), which implicitly uses cut-off sampling, without paying too much attention to methodological implications: "The sampling frame for the Canadian Monthly Survey of Manufacturing (MSM) is determined from the target population after subtracting establishments that represent the bottom 2% of the total manufacturing shipments estimate for each province. These establishments were excluded from the frame so that the sample size could be reduced without significantly affecting quality". Similar procedures are also employed in surveys performed by other National Statistical Offices: cut-off sampling is widely used but methodological aspects are not documented.

Two exceptions are the book by Särndal *et al.* (1992, pagg. 531-533), who are mostly negative, and the paper by Haan *et al.* (1999), who present successful applications of cut-off sampling in the field of consumer price indexes.

Finally, Elisson and Elvers (2001) performed a univariate analysis which compares cut-off sampling with simple stratified sampling. They conclude that cut-off sampling is worth more consideration and suggest to use it in applications; however, they find that the dimensional variable which determines the cut-off threshold has a relevant impact on the results, so that they stress that great care must be employed in choosing this variable. Moreover, they point out the need for an appropriate model to estimate the fraction of population excluded from the sample.

In any case, it is worth mentioning the practical advantages of cut-off sampling as concerns the costs of a survey:

(i) building and updating a sampling frame for small business units could be too costly, considering that the gain in efficiency of the estimators would probably be small;

(ii) excluding the units of the population which give little contribution to the aggregates to be estimated usually implies a large decrease of the number of units which have to be surveyed in order to get a predefined accuracy level of the estimates.

(iii) putting a constraint to the frame population and, as a consequence, to the sample, allows to reduce the problem of empty strata which mainly affects the smallest firms.

As of this issue, it is worth stressing that several empirical analyses showed that some difficulties, such as the non-response rate, the natimortality of the economic units and the errors of under- or over-coverage of the frame, become more relevant as the size of the units gets small.

Given that practitioners are in favor of such partitions of the population and there are technical reasons which justify their use, the basic question is: is it possible to consider cut-off sampling as a valid sampling scheme? If the answer is positive, the issue is to define a statistical framework for cut-off sampling.

In this work we try to develop an easily implementable solution to the problem of the construction of the three strata U_C, U_S and U_E in a multipurpose and multivariate setup. In other words, similarly to what happens in practical applications, we assume to be interested in surveys with more than one target variable, using auxiliary information contained in multiple variables.

The structure of the paper is as follows. In section 2 we will define an estimation model which assumes, for each variable, the weight of the units excluded from the analysis to be known and constant; however, this hypothesis is not, in general, under the control of the investigator, so that this estimator is biased, and we will have to find the bias and the mean squared error of the estimator. The model will be developed both for the estimation of a total and for the estimation of a ratio of totals. Section 3 will be devoted to the derivation of the sample size for the cut-off scheme, both when estimating a total and when estimating a ratio, focusing on its optimization and, consequently, on the construction of the optimal design. The problem will be tackled by defining the sample size as a function of the partition U_C, U_S and U_E determined on the basis of multivariate auxiliary information which will be assumed to be known for the whole population. Considering the combinatorial nature of this problem, we will use the theory of stochastic relaxation and, in particular, the Simulated Annealing (SA) algorithm. In section 4 we will show some empirical evidence about the bias of the estimator when using data from surveys concerning slaughtering firms in Italy. In the same section we will present the main results of the application of the sampling scheme to this dataset. Finally, section 5 shall conclude the paper and point out some open problems.

2. An estimator for cut-off sampling schemes

The problem of stratifying in two strata (take-all and take-some) and finding the census threshold was first treated by Dalenius (1952) and Glasser (1962). The first author has found the census threshold as a function of the mean, the sampling weights and the variance of the population. Glasser (1962) derived the value of the threshold under the hypothesis of

sampling without replacement a sample of size n from a population of N units. Hidiroglou (1986) reconsidered this problem and provided both exact and approximate solutions under a more realistic hypothesis: he finds the census threshold when a level of precision concerning the mean squared error of the total is desired, without assuming a predefined sample size n. It is worth noticing that he considers a case with only a take-all and a take-some stratum, so that he develops a method for finding a "census threshold" (erroneously defined "cut-off threshold" in the paper). However, all these authors limit their attention to a monopurpose and univariate setup.

Hidiroglou's approach will be followed in the present paper as well, but here we will stratify the target population by means of a criterion which defines the belonging of each observation to one of the three strata in a multipurpose and multivariate framework. The solution of the problem is based on the identification of appropriate estimators for the quantities in table 1.

Table 1. Estimators and error measures; $b(\cdot)$ is the bias function, f and g are functions which shall be defined in the following

Stratum	U_C	U_S	U_E
Estimator	$\hat{t}_C$	$\hat{t}_S$	$f(\hat{t}_C, \hat{t}_S)$
Estimator MSE	0	$\mathrm{var}(\hat{t}_S)$	$g(\mathrm{var}(\hat{t}_S) + b^2(\hat{t}_E))$

2.1. Estimating a total

We start by considering the estimator of the total $\hat{t}_{y_j}$ of the j-th surveyed variable ($j = 1, \ldots, J$). This estimator is the sum of three independent components, corresponding respectively to the take-all, take-some and take-nothing strata. Thus, omitting for simplicity the index of the variables (the same way of reasoning can be applied to all the J variables once the belonging criterion mentioned above has been determined), we can write $\hat{t}_y = \hat{t}_C + \hat{t}_S + \hat{t}_E$. As for the take-all stratum, it is clear that $\hat{t}_C = \sum_{k \in U_C} y_k$. In the take-some stratum, we use the classical $\pi-$estimator of the total $t_S = \sum_{k \in U_S} y_k$:

$$\hat{t}_{\pi S} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k, \tag{1}$$

which is the expansion formula known in the literature as Horvitz-Thompson estimator (Horvitz and Thompson 1952). In (1), the π_k's are the inclusion probabilities, which are assumed to be strictly positive; the same condition holds for the second-order probabilities π_{kl}, which are necessary for the computation of the variance of the estimator. The quantities

$d_k = 1/\pi_k$ are the direct weights of each unit $k \in s$, namely the original weights resulting from the sampling scheme.

The sample s is a probabilistic sample drawn from the subpopulation U_S; in the following we will always assume that it is a simple random sample from U_S. According to the setup of our problem, the Hidiroglou-type estimator $\hat{t}_C + \hat{t}_S = \sum_{k \in U_C} y_k + \sum_{k \in s} d_k y_k$ has to be augmented by a model-based component which takes into account the discarded fraction of the population, U_E. As concerns this issue, we can write

$$t_E = (t_C + t_S)\delta, \tag{2}$$

i.e., the total of the discarded population is a fraction of $t_C + t_S$. In (2) the quantity δ, which is usually unknown, can be evaluated by means of external sources (i.e., the auxiliary variables $\boldsymbol{x}$); thus

$$\tilde{\delta} = \frac{\sum_{k \in U_E} x_k}{\sum_{k \in U_C} x_k + \sum_{k \in U_S} x_k}. \tag{3}$$

For notational simplicity and without loss of generality, in the following we will always assume that each auxiliary variable is the lagged target variable (in most cases, as well as in the present application, it is the target variable as known from the last census): $x_k = y_{k,t-1}$. Using these hypotheses we obtain the following identity:

$$\hat{t}_y = \hat{t}_C + \hat{t}_S + \hat{t}_E = (1 + \tilde{\delta})(\hat{t}_C + \hat{t}_S) = (1 + \tilde{\delta}) \left(\sum_{k \in U_C} y_k + \sum_{k \in s} d_k y_k \right). \tag{4}$$

The hypotheses introduced to obtain (4) are slightly different from Särndal *et al.* (1992, pag. 532), who use a ratio estimator in the domain S as a "compensation" for the fraction of population discarded. As we are concerned with a sampling design, in this paper we find it more convenient to employ, as a starting point for the part of the population to be sampled, the "neutral" Horvitz-Thompson estimator. However, it is worth pointing out that there is no reason which prevents us from implementing, in the estimation procedure, a second step: we could indeed use the auxiliary information *ex post*, in order to correct $\hat{t}_C$ and $\hat{t}_S$ either by means of a ratio estimator or by means of a more general approach to the use of auxiliary information such as the so called calibration estimators (Deville and Särndal 1992). In addition to several desirable properties, calibration estimators possess a very important feature, namely they reduce the bias arising from total nonresponses, which would also appear when enumerating completely the subpopulation U_C.

It is well known (see, for example, Särndal *et al.* 1992, pag. 531) that cut-off sampling produces biased estimators. Using (4) and the independence of the three strata U_C, U_S and

U_E, the Mean Squared Error of $\hat{t}_y$ is given by:

$$\begin{aligned} MSE(\hat{t}_y) &= \mathrm{var}(\hat{t}_y) + b^2(\hat{t}_y) = \mathrm{var}(\hat{t}_C + \hat{t}_S + \hat{t}_E) + b^2(\hat{t}_y) = \\ &= \mathrm{var}[(1+\tilde{\delta})(\hat{t}_C + \hat{t}_S)] + b^2(\hat{t}_y) = (1+\tilde{\delta})^2 \mathrm{var}(\hat{t}_C + \hat{t}_S) + b^2(\hat{t}_y) = \\ &= (1+\tilde{\delta})^2 \mathrm{var}(\hat{t}_S) + b^2(\hat{t}_y) = (1+\tilde{\delta})^2 \mathrm{var}(\hat{t}_S) + b^2(\hat{t}_E). \end{aligned} \tag{5}$$

In (5) we put $b(\hat{t}_y) = b(\hat{t}_E)$ to stress that the bias, which represents the price to pay for discarding part of the population, only depends on excluded strata. It is indeed clear that $\tilde{\delta} \in I\!R^+$ in (4) introduces a bias because the true ratio δ of the discarded to the completely enumerated and sampled population is unknown and different from the estimated value $\tilde{\delta}$ which is used in the current survey.

It is therefore crucial to concentrate on the bias $b(\hat{t}_E)$. It is not difficult to see that:

$$\begin{aligned} b(\hat{t}_E) &= E(\hat{t}_y) - t_y = E(\hat{t}_C + \hat{t}_S + \hat{t}_E) - t_y = \\ &= \sum_{k \in U_C} y_k + \sum_{k \in U_S} y_k + E[\tilde{\delta}(\hat{t}_C + \hat{t}_S)] - t_y = \\ &= \tilde{\delta}(t_C + t_S) - t_E. \end{aligned} \tag{6}$$

Putting $t_E = \delta(t_C + t_S)$, (6) can be conveniently rewritten as follows:

$$b(\hat{t}_y) = (\tilde{\delta} - \delta)(t_C + t_S). \tag{7}$$

From (7) it appears that the source of the bias of the estimator (4) is the mismatch between the numerical value $\tilde{\delta}$ used in the survey and the true value δ; in particular, the magnitude of the bias is determined by the difference $|\tilde{\delta} - \delta|$.

As will become clearer in the next section, (7) is a fundamental ingredient of the sample design proposed here. In section 4 we will show some empirical evidence concerning the functional form of the bias.

2.2. Estimating a ratio

Suppose now that the aim of the investigator consists in estimating not just a total but a ratio of two unknown totals: $R = t_y/t_z = \sum_U y_k / \sum_U z_k$. The usual estimator (Särndal *et al.* 1992, pag. 176-181) is a non-linear function of the two random variables $\hat{t}_{y\pi}, \hat{t}_{z\pi}$:

$$\hat{R} = f(\hat{t}_{y\pi}, \hat{t}_{z\pi}) = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}.$$

In some applications, including conjunctural surveys, we are interested in the estimation of the ratio $R = t_{y,t}/t_{y,t-1}$ or in the variation $R - 1$.

When, as is the case in our setup, the sampling scheme used in U_S is simple random sampling without replacement, the following identities hold:

$$\begin{aligned} n &= fN; \\ \hat{t}_{y\pi} &= \sum_s \frac{y_k}{\pi_k} = \sum_s \frac{N}{n} y_k = N \sum_s \frac{y_k}{n} = N\bar{y}_s; \\ \hat{t}_{z\pi} &= N\bar{z}_s; \\ \hat{R} &= \frac{\bar{y}_s}{\bar{z}_s}, \end{aligned}$$

where $f = n/N$ is the sampling fraction. The first-order Taylor expansion of $\hat{R}$ (Särndal *et al.* 1992, sect. 5.5) gives the following result:

$$\begin{aligned} \hat{R} \approx \hat{R}_0 &= R + \frac{1}{t_z} \sum_s \frac{y_k - Rz_k}{\pi_k} = R + \frac{1}{t_z}(\hat{t}_{y\pi} - R\hat{t}_{z\pi}) = \\ &= R + \frac{1}{t_z} \frac{N}{n} \sum_s (y_k - Rz_k) = \\ &= R + \frac{\bar{y}_s - R\bar{z}_s}{\bar{z}_U}, \end{aligned} \tag{8}$$

where $\bar{z}_U = t_z/N$. In our setup the population is enumerated completely at the time preceding $t-1$, that is $t-2$. We assume that at time $t-2$ the cut-off design has been implemented, so that a global sample of business units sampled and enumerated completely for the survey at times $t-1$ and t is available. Thus

$$\hat{R}_0 = R + \frac{\bar{y}_{s,t} - R\bar{y}_{s,t-1}}{\bar{y}_{U,t-1}}.$$

This estimator is approximately unbiased:

$$E(\hat{R}) \approx E(\hat{R}_0) = R.$$

As we take a linear (first-order) approximation of $\hat{R}$, the approximation error is given by the fact that we ignore the terms of order larger than one in the Taylor expansion (8); in other words, the approximation error is given by the "nonlinear component" of $\hat{R}$.

Following the same way of reasoning of the preceding subsection we get:

$$\begin{aligned} \hat{R}_{0y} &= \hat{R}_{0C} + \hat{R}_{0S} + \hat{R}_{0E} = (1+\tilde{\delta})(\hat{R}_{0C} + \hat{R}_{0S}); \\ MSE(R_{0y}) &= \mathrm{var}(R_{0y}) + b^2(\hat{R}_{0y}) = (1+\tilde{\delta})^2[\mathrm{var}(\hat{R}_{0S}) + b^2(\hat{R}_{0E})]; \\ b(\hat{R}_{0E}) &= (\tilde{\delta} - \delta)(R_C + R_S). \end{aligned}$$

As for the computation of $\mathrm{var}(\hat{R}_{0S})$, we use the following approximation, again derived by means of a Taylor expansion (Särndal 1992, pag. 178), and only valid in the case of simple

random sampling without replacement in U_S:

$$\text{var}(\hat{R}) \approx \text{var}(\hat{R}_0) = \frac{1}{t_z^2}[\text{var}(\hat{t}_{y\pi}) + R^2\text{var}(\hat{t}_{z\pi}) - 2R\text{cov}(\hat{t}_{y\pi}, \hat{t}_{z\pi})]. \tag{9}$$

With a more general notation, (9) can be rewritten as

$$\text{var}(\hat{R}) \approx \text{var}(\hat{R}_0) = \frac{1}{\bar{z}^2}\frac{1-f}{n}(S_{yU}^2 + R^2 S_{zU}^2 - 2RS_{yzU}),$$

where S_{yzU} is the covariance between y and z in the population.

Following Hidiroglou's (1986) terminology, the term $(1-f)/n$ in the take-some stratum takes the form:

$$\frac{1-f}{n} = \frac{N-n(t)}{(N-t)(n(t)-t)}. \tag{10}$$

Thus, in our setup, $(1-f)/n$ can be rewritten as:

$$\frac{1-f}{n} = \frac{N-n}{(N-N_C-N_E)(n-n_C)}.$$

The variance of $\hat{R}_{0S}$ can be put in the form

$$\text{var}(\hat{R}_{0S}) = \frac{1}{\bar{y}_{U_S,t-1}^2}\frac{N-n}{(N-N_C-N_E)(n-n_C)}(S_{U_S,t}^2 + R_S^2 S_{U_S,t-1}^2 - 2S_{U_S,y_{t-1}y_t}). \tag{11}$$

With no additional information (such as, for example, a variance trend which could possibly be extrapolated), when implementing the sampling design the only reasonable assumption is that

$$S_{U_S,t-2}^2 = S_{U_S,t-1}^2 = S_{U_S,t}^2 = S_{U_S}^2,$$

so that $S_{U_S,y_{t-1}y_t} = \rho_{U_S,y_{t-1}y_t}S_{U_S}^2$, where $\rho_{U_S,y_{t-1}y_t}$ is the correlation coefficient of the variables y_t and y_{t-1} in the population. Plugging this result into (11) we finally have

$$\text{var}(\hat{R}_{0S}) = \frac{1}{\bar{y}_{U_S,t-1}^2}\frac{N-n}{(N-N_C-N_E)(n-n_C)}(1 + R_S^2 - 2\rho_{U_S,y_{t-1}y_t}R_S)S_{U_S}^2.$$

In the next section we will use this variance to determine the optimal sample size.

3. The sample size for cut-off and optimal designs

3.1. Sample size when estimating a total

In the preceding section we showed that the MSE of the estimator of the total $\hat{t}_y$ for cut-off designs is equal to $MSE(\hat{t}_y) = (1+\tilde{\delta})^2\text{var}(\hat{t}_{\pi S}) + b^2(\hat{t}_y)$, where the first term is the variance of the Horvitz-Thompson estimator used for estimating the total of the target variables in the subpopulation U_S.

The well-known expression for this variance in simple random sampling without replacement (Särndal *et al.* 1992, pag. 46) is given by

$$\mathrm{var}(\hat{t}_\pi) = N^2 \frac{1-f}{n} S^2, \tag{12}$$

where S^2 is the variance of the target variable. However, in our setup this formula needs to be modified: the Horvitz-Thompson estimator is indeed only used in U_S, so that

$$\mathrm{var}(\hat{t}_{\pi S}) = N^2 \frac{(N - N_C - N_E)(N-n)}{n - n_C} S^2. \tag{13}$$

In (13) the variance S^2 is equal to

$$S^2_{U_S} = \frac{1}{N - N_C - N_E - 1} \sum_{k \in U_S} (y_k - \mu)^2,$$

where $\mu = \mu_{U_S} = (1/(N - N_C - N_E)) \sum_{k \in U_S} y_k$.

In applications, the MSE is usually required to satisfy the following equality:

$$MSE(\hat{t}_y) = c^2 t_y^2, \tag{14}$$

where c is the desired level of precision c for the estimator of the total. If we substitute for $MSE(\hat{t}_y)$ in (14) the second term on the right hand side of (5) we get:

$$(1 + \tilde{\delta})^2 \mathrm{var}(\hat{t}_{\pi S}) + b^2(\hat{t}_y) = c^2 t_y^2,$$

from which we easily derive the variance of the estimator:

$$\mathrm{var}(\hat{t}_{\pi S}) = \frac{(N - N_C - N_E)(N-n)}{n - n_C} S^2 = \frac{c^2 t_y^2 - b^2(\hat{t}_y)}{(1 + \tilde{\delta}^2)}. \tag{15}$$

We now focus on expression (15) in order to derive the total sample size. Here, the size is defined to be "total" because it includes both the size of the stratum completely enumerated and of the simple random sample without replacement from the stratum U_S. In the following it obviously holds that $n_C = N_C = N - N_S - N_E$; for notational simplicity, we first put $\psi = [c^2 t_y^2 - b^2(\hat{t}_y)]/(1 + \tilde{\delta})^2$. We have

$$\frac{(N - N_C - N_E)(N-n)}{n - n_C} S^2 = \psi,$$

from which we get

$$(N - N_C - N_E) N S^2 + n_C \psi = n(N - N_C - N_E) S^2 + n\psi.$$

Solving with respect to n we obtain

$$n = \frac{(N - N_C - N_E) N S^2 + n_C \psi}{\psi + (N - N_C - N_E) S^2}. \tag{16}$$

With some more algebra it is possible to obtain the following result, which is preferable from a computational point of view:

$$n = N - \frac{1}{\frac{1}{N_S+N_E} + \frac{N_S}{N_S+N_E}\frac{S^2}{\psi}}. \tag{17}$$

3.2. Sample size when estimating a ratio

As in the preceding subsection, we start with a predetermined level of precision c concerning the estimator of the ratio:

$$MSE(\hat{R}_{0y}) = c^2 R_y^2. \tag{18}$$

Following the same way of reasoning used before, we can rewrite (18) as

$$(1+\tilde{\delta})^2 \mathrm{var}(\hat{R}_{0S}) + b^2(\hat{R}_{0E}) = c^2 R_y^2,$$

so that

$$\mathrm{var}(\hat{R}_{0S}) = \frac{c^2 R_y^2 - b^2(\hat{R}_{0E})}{(1+\delta)^2} \stackrel{def}{=} \psi_1.$$

Putting $\psi_2 = 1 + R_S^2 - 2\rho_{U_S, y_{t-1} y_t} R_S$, we get

$$\frac{N_S^2}{\hat{t}_{S,t-1}^2} \frac{N-n}{(N-N_C-N_E)(n-n_C)} \psi_2 S_{U_S}^2 = \psi_1.$$

Some straightforward algebra gives:

$$n\left[\psi_1 + \psi_2 S_{U_S}^2 \frac{N-N_C-N_E}{\hat{t}_{S,t-1}^2}\right] = \frac{N-N_C-N_E}{\hat{t}_{S,t-1}^2} N \psi_2 S_{U_S}^2 + n_C \psi_1,$$

from which we obtain the optimal sample size:

$$n = \frac{N(N-N_C-N_E)\psi_2 S_{U_S}^2 + n_C \hat{t}_{S,t-1}^2 \psi_1}{\hat{t}_{S,t-1}^2 \psi_1 + \psi_2 S_{U_S}^2 (N-N_C-N_E)}.$$

Finally, it is not difficult to show that the sample size n can be rewritten as follows:

$$n = N - \frac{1}{\frac{1}{N_S+N_E} + \frac{N_S}{N_S+N_E}\frac{\psi_2}{\psi_1}\frac{S_{U_S}^2}{\hat{t}_{S,t-1}^2}}.$$

3.3. Optimal partition

In (17) the sample size n depends on c, which is chosen *a priori* by the researcher, on the bias $b(\hat{t}_E)$, on the total t_y and on the partition in the three strata. Notice that the latter determines four additional quantities, namely $\tilde{\delta}$, N_S, N_E and S^2. Thus, if we denote with $\Phi = \{k_1, k_2, \ldots, k_N\}$ ($k_i \in \{C, S, E\}$) the generic element of the set Θ of the possible

partitions of the population (whose cardinality is equal to 3^N), we conclude that n is a function of Φ and write

$$n = n(\Phi), \tag{19}$$

because all the other quantities listed above are either chosen by the researcher or computed using the auxiliary variables once a partition has been determined.

At this point it is quite clear that the problem consists in finding the partition Φ^* which minimizes (19) given the desired level of precision c. In particular, as our aim is the estimation of the totals t_{y_j} of J variables by means of the same number J of auxiliary variables (see section 2.1), the optimal sample size can be defined as follows:

$$n(\Phi^*) = \min \left\{ \max_{j=1,\ldots,J} n_j(\Phi) \right\}. \tag{20}$$

The term $\max_{j=1,\ldots,J} n_j(\Phi)$ in (20) means that the optimization concerns, at each iteration, the largest of the sample sizes n_j corresponding to each auxiliary variable. (20) is the formalization of a combinatorial optimization problem. The simulated annealing (Metropolis *et al.* 1953, Kirkpatrick 1983, Geman and Geman 1984) is probably the best suited method for solving (20). This algorithm, which belongs to the family of stochastic relaxation algorithms, enjoys several desirable properties (see Casella and Robert 1999, sect. 5.2.3, for a review); its implementation to the problem at hand can be summarized as follows.

(a) Choose an initial temperature T_0.

(b) Stratify the population by means of a random uniform partition Φ_0, that is, assign to each of the N units of the population a label ϕ from the set $\{C, S, E\}$, where $P(\phi = C) = P(\phi = S) = P(\phi = E) = 1/3$. Let $\phi_i^{(0)}$ $(i = 1, \ldots, N)$ be these labels.

(c) Visit the i-th unit of the population and put $\phi_i^{(1)} = \xi$, where ξ is a label drawn with uniform probability from the set $\{C, S, E\}$ and is the update of the label assigned to the i-th unit at the 0-th iteration. Obviously, $\phi_j^{(1)} = \phi_j^{(0)}$ $\forall j \neq i$, so that the vector of labels $\boldsymbol{\phi}^{(1)}$ at the first iteration differs from $\boldsymbol{\phi}^{(0)}$ at most by one element.

(d) Let $\Delta^{(1)} = n(\Phi^{(1)}) - n(\Phi^{(0)})$. If $\Delta < 0$, put $\phi_i^{(1)} = \xi$; otherwise, put $\phi_i^{(1)} = \xi$ with probability $\exp\{\Delta^{(1)}/T_0\}$ or $\phi_i^{(1)} = \phi_i^{(0)}$ with probability $1 - \exp\{\Delta^{(1)}/T_0\}$.

(e) Repeat step 3. and 4. $(N_{sub} \times N)$ times, where N_{sub} is the number of sub-iterations for each temperature T.

(f) Replace T_0 with $T_1 = f(T_0)$, where $f(\cdot)$ is a decreasing function which satisfies the conditions of the *annealing theorem* (Geman and Geman 1984). The function originally proposed by Geman and Geman (1984) was $T_{t+1} = f(T_t) = (\log(1+t)/\log(2+t))T_t$; here we follow Sebastiani (2003) and use the so-called geometric temperature schedule $T_{t+1} = f(T_t) = \rho T_t$, with $\rho \in (0, 1)$. The choice of f in applications has been the object of a lot of interest and some controversial in the literature: see Ripley (1988),

Stander and Silverman (1994), Winkler (1995) and Casella and Robert (1999, pag. 201), and the references therein. As for the numerical value of ρ, it is well known that it has to be "large" enough to avoid a too rapid decrease of the temperature and "small" enough to keep the computation time reasonably short. We performed several experiments and found that $\rho = 0.98$ guarantees the best compromise.

(g) Repeat steps 3-6 until some convergence criterion is satisfied. We found it convenient to stop the algorithm the first time that one of the following conditions is satisfied: (i) in two successive iterations no labels are switched; (ii) $n_{iter} = 300$ iterations are reached.

Notice that at step 7. the t-th iteration is just obtained by replacing (0) with (t) and (1) with $(t+1)$ in steps 3-6 above.

At convergence, the algorithm determines the optimal partition Φ^*, which minimizes the total sample size n for a given precision level c.

4. A case study: the slaughtering monthly survey

In this section we will find the optimal design, according to the cut-off methodology developed so far, for the red meat slaughtering monthly survey performed by ISTAT (Italian National Institute of Statistics). This survey foresees a stratified sampling, with a stratification by kind of slaughter-houses and geographical division, for a total of 5 strata, two of which with geographical references. Strata are the following: stratum 1 (always totally observed), consisting of private with European Economic Community (EEC) stamp slaughter-houses in the geographical division 1 or 2; stratum 2: consisting of private with EEC stamp slaughter-houses in the geographical division 3, 4 or 5; stratum 3: private with low capacity slaughter-houses (apart from geographical division); stratum 4: private in derogation, public with EEC stamp and public in derogation (apart from geographical division) slaughter-houses; stratum 5: public with low capacity slaughter-houses. Two dimensional criteria that assign to stratum 1 those enterprises with more than 10000 sheep and goats or more than 50000 pigs slaughterings act in the stratification too. On the average the sample is of about 460 units for a population of 2211 units with the desired level of precision c set to 5%.

Thus, our frame contains $N = 2211$ slaughter-houses for which we know four variables enumerated completely in 1999, 2000 and 2001 (consider indeed that this census is performed every year): they are respectively the total number of slaughtered (i) cattle, (ii) pigs, (iii) sheep and goats and (iv) equines. We will first consider the complete dataset (for each of the three years) in order to assess the behavior of the bias $b(\hat{t}_E)$ and, in particular, to look for possible regularities. Recalling that δ is defined as the ratio of the number of population

units discarded to the number of population units sampled and enumerated completely, it is indeed necessary to know, in order to evaluate empirically the bias, the complete list of the lagged auxiliary variables.

The cut-off design proposed in this paper will then be implemented, with the aim of setting up a monthly survey on slaughtering for the year 2002, using as auxiliary variables only the data enumerated completely in 2001.

We start with a brief description of the archive at hand. The scatterplots of all the pairs of the four variables in 2001 are shown in the above-diagonal graphs of figure 1; the graphs below the main diagonal are the scatterplots of the logs of the same pairs of variables. The same graphs for the years 1999 and 2000 are essentially identical and therefore are not reported here.

The main evidence is that variables are essentially independent or, in some cases, negatively correlated. Moreover, it is clear that slaughter-houses are strongly specialized and that most firms are small (see the histograms on the main diagonal).

4.1. Bias assessment

In order to implement the design developed in the preceding sections, it is crucial to analyze the bias $b(\hat{t}_E)$ of the estimator $\hat{t}_y$ given by (4), because the algorithm described in the preceding section requires as an input a starting value for $b(\hat{t}_E)$. We solved this problem with the help of both empirical evidence concerning real data and simulations of auxiliary variables with different skewness.

As for the real data, figure 2 shows some very interesting results. Here, we plotted the absolute value of the bias $b(\hat{t}_{E_i})$, where the quantity $\hat{t}_{E_i}$ is defined as the total of the discarded population observed in 1999 and 2000, which in turn is given by the i smallest observations of the population. In other words, the i-th point of the graph is the absolute value of the bias corresponding to $\hat{t}_{E_i}$, where E_i contains the i smallest observations of the population.

The procedure used to estimate the bias works as follows. If a complete enumeration of both the auxiliary variable $\boldsymbol{x}$ and the objective $\boldsymbol{y}$ (usually they are the same variable relevant to two different periods) are available, they can be ordered on the basis of the values of $\boldsymbol{x}$:

$$x_{(1)}, \ldots, x_{(N)},$$
$$y_{(1)}, \ldots, y_{(N)},$$

where the (i) codes are such that $x_{(i)} \leq x_{(i+1)}$ for $i = 1, 2, \ldots, N-1$. Let now $C_{x,(i)}$ and

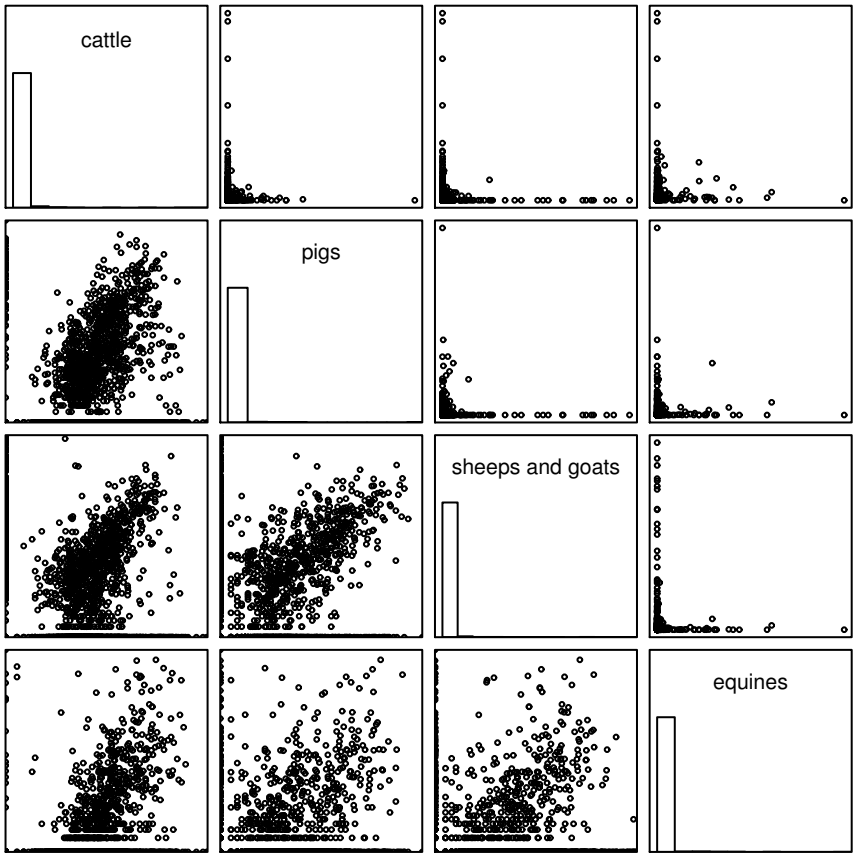

Fig. 1. Scatterplots of the data.

$C_{y,(i)}$ be the respective cumulative sums:

$$C_{x,(i)} = \sum_{j=1}^{i} x_{(j)},$$
$$C_{x,(i)} = \sum_{j=1}^{i} x_{(j)},$$

and $O_{x,(i)}$ and $O_{y,(i)}$ be the corresponding countercumulative sums:

$$O_{x,(i)} = t_x - C_{x,(i)} = \sum_{j=i+1}^{N} x_{(j)},$$
$$O_{y,(i)} = t_y - C_{y,(i)} = \sum_{j=i+1}^{N} y_{(j)}.$$

Thus, if i is used as a threshold, according to (6), the absolute value of the bias obtained by using the estimator (4) can be written as:

$$|b_i| = \left| C_{x,(i)} \frac{O_{y,(i)}}{O_{x,(i)}} - C_{y,(i)} \right|,$$

where the excluded part of the population is defined as $E_i = \{1, 2, \ldots, i\}$.

The $|b_i|$ can be used either directly in the optimization algorithm or modelled to simplify calculations and to obtain more stable results, i.e. not depending on particular discontinuities in the frame data. In our experiment we found good fits for the simple linear regression model:

$$|b_i| = \alpha + \beta C_{x,(i)} + \epsilon_i. \tag{21}$$

The fit becomes better if the tails of the ordered distributions are dropped out from the analysis, but this is not a problem because in practical applications a threshold is usually neither a very small nor a very large value.

The four graphs in figure 2, corresponding to each auxiliary variable, have been obtained using respectively the complete 1999 and 2000 frame as a basis for the construction of the cut-off design in 2001.

As expected, a larger temporal lag of the auxiliary information causes a significant modification of the bias: the bias for 2001 is indeed always larger than the bias for 2000. Moreover, from the graphs it appears that the function f which formalizes the relationship between $|\hat{b}(t_{E_i})|$ and $C_{x,(i)}$ is well fitted by the linear model (21).

We now apply our cut-off procedure in order to re-design the ISTAT red meat slaughtering monthly survey; to this aim we will use, as auxiliary information, the aforementioned frame for the year 2001. Thus, at each iteration of the algorithm and for each auxiliary

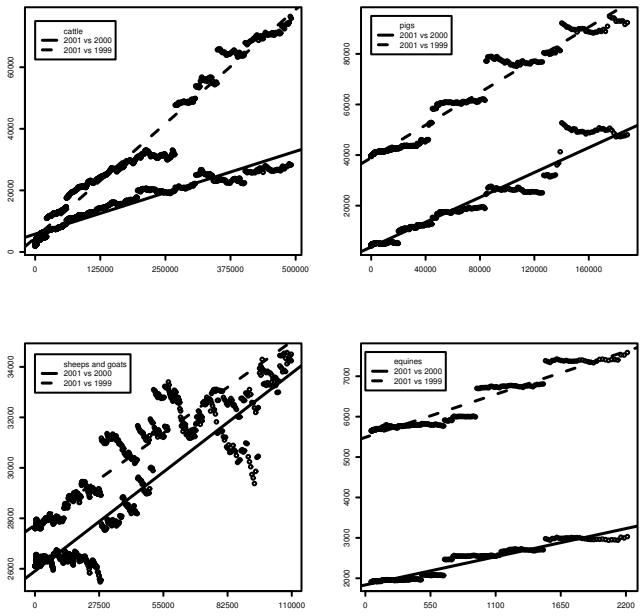

Fig. 2. The relationship between $|\hat{b}(t_{E_i})|$ and $C_{x,(i)}$.

variable, we substitute to the bias $b(\hat{t}_E)$ which appears in the expression (16) for $n(\Phi^{(t)})$, an estimate obtained via linear regression, i.e.:

$$|\hat{b}_i(t_{E_i})|_{j,2001} = \hat{\alpha} + \hat{\beta} C_{j,2000(i)}, \qquad j = 1, \ldots, 4, \tag{22}$$

$$|\hat{b}_i(t_{E_i})|_{j,2001} = \hat{\alpha} + \hat{\beta} C_{j,1999(i)}, \qquad j = 1, \ldots, 4. \tag{23}$$

Equations (22) and (23) actually give an estimate of the absolute value of the bias, but this is not relevant because (16) only uses the square of this estimate. Detailed results are displayed in tables 2 to 5 respectively for cattle, pigs, sheep and goats, equines.

Table 2. Cattle: estimates, standard errors, t-statistics and p-values for (22) and (23).

	2001 vs 1999 ($R^2 = 0.9494$)			2001 vs 2000 ($R^2 = 0.9662$)		
	estimate	t-stat	p-value	estimate	t-stat	p-value
α	5810(47.75)	121.7	< 0.0001	4621(106.6)	43.7	< 0.0001
β	0.0538(0.0003)	175.9	< 0.0001	0.1482(0.0007)	217.25	< 0.0001

Table 3. Pigs: estimates, standard errors, t-statistics and p-values for (22) and (23).

	2001 vs 1999 ($R^2 = 0.9541$)			2001 vs 2000 ($R^2 = 0.96$)		
	estimate	t-stat	p-value	estimate	t-stat	p-value
α	3664(96.48)	37.98	< 0.0001	38930(118.1)	329.6	< 0.0001
β	0.2455(0.0019)	130.0	< 0.0001	0.3230(0.0023)	139.7	< 0.0001

Table 4. Sheep and goats: estimates, standard errors, t-statistics and p-values for (22) and (23).

	2001 vs 1999 ($R^2 = 0.8517$)			2001 vs 2000 ($R^2 = 0.9158$)		
	estimate	t-stat	p-value	estimate	t-stat	p-value
α	25910(39.37)	658.09	< 0.0001	27740(26.53)	1045.31	< 0.0001
β	0.0713(0.0011)	66.88	< 0.0001	0.0661(0.0007)	92.04	< 0.0001

Table 5. Equines: estimates, standard errors, t-statistics and p-values for (22) and (23).

	2001 vs 1999 ($R^2 = 0.9286$)			2001 vs 2000 ($R^2 = 0.9169$)		
	estimate	t-stat	p-value	estimate	t-stat	p-value
α	1834(9.191)	199.56	< 0.0001	5488(15.21)	360.8	< 0.0001
β	0.6344(0.0107)	59.15	< 0.0001	0.9674(0.0177)	54.5	< 0.0001

The fit is extremely good in all cases; in particular, the values of the R^2 goodness-of-fit statistics are always large, which is not surprising if we consider that the variables used in the regression are cumulative sums.

4.2. Sampling design

Let's now finally turn to the results of the implementation of the cut-off design. Figure 3 shows the total optimal sample size as a function of the number of iterations of the simulated annealing.

It is immediately evident that the "largest decrease" in the sample size takes place in the first few iterations; the remaining iterations seem to provide us with just an adjustment towards the global optimum. More precisely (see figure 4), starting from the third iteration, the algorithm just moves some observations from U_E to U_S; to these label-switching operations correspond very small decreases of the total sample size.

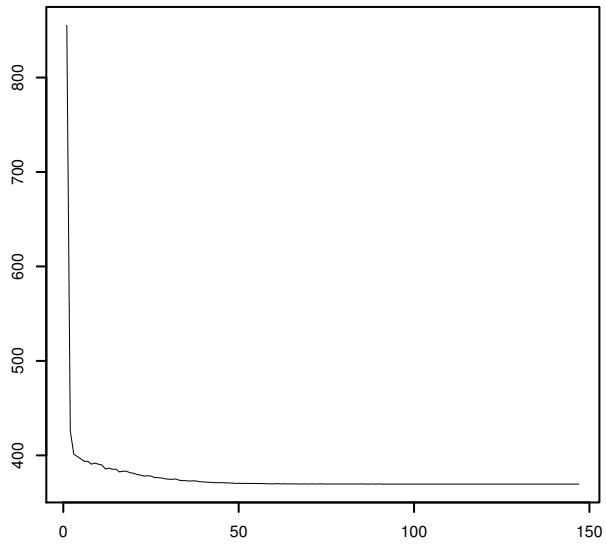

Fig. 3. Total sample size $n = N_C + n_S$ as a function of the SA iterations.

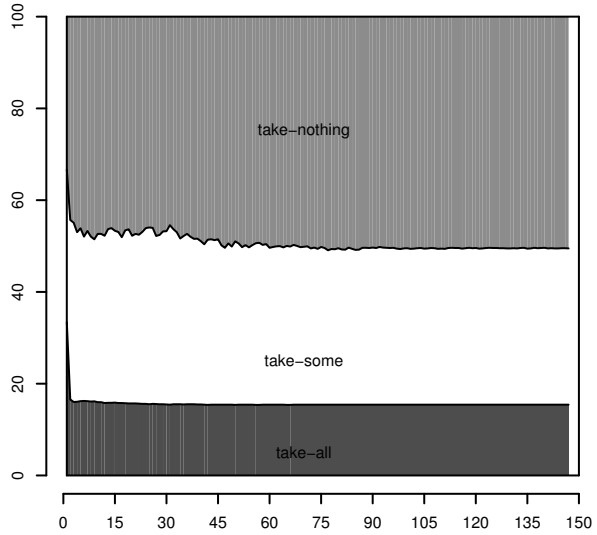

Fig. 4. Percentage composition of the strata U_C, U_S and U_E as a function of the SA iterations.

Table 6 gives some details about the implementation of the algorithm. The quantity N_S is the size of the stratum U_S; the number of units actually sampled from this stratum can be computed as $n - N_C$; for example, at the 147-th iteration (namely when the algorithm converges) we sample $n - N_C \approx 370 - 341 = 29$ units.

Table 6. Results of the cut-off sampling as a function of selected iterations of SA.

Iter	n	N_C	N_S	N_E	# changes
1	855.30	738	736	737	-
2	425.97	366	865	980	3047
3	401.10	354	864	993	2233
4	398.80	355	817	1039	2039
5	396.30	358	833	1020	1941
6	393.70	359	792	1060	1952
7	393.70	358	820	1033	1861
8	390.70	356	795	1060	1873
9	391.80	357	781	1073	1825
10	390.70	353	811	1047	1797
20	381.10	347	808	1056	1385
30	374.80	342	835	1034	889
40	371.90	342	787	1082	548
50	370.44	340	788	1083	309
60	369.90	341	756	1114	174
100	369.60	341	754	1116	19
147	369.60	341	753	1117	2

The sampling scheme developed in this paper produces the partition of the population shown in figure 5.

Each subplot of this figure gives the scatterplot of the fourth roots of the auxiliary variables. The stratification is very clear-cut, with two strata (U_C and U_E) whose sizes are much larger than U_S. The take-some stratum is nested into the take-nothing stratum, with a sampling fraction equal to 4%: this means that in our application the sampling scheme is very similar to a take-all/take-nothing design. According to the theoretical results derived in the preceding section, such a small sampling fraction was indeed expected: considering the large concentration of the population, stratum U_S contains mostly the firms for which the values of all the four auxiliary variables are different from zero, namely the least specialized ones.

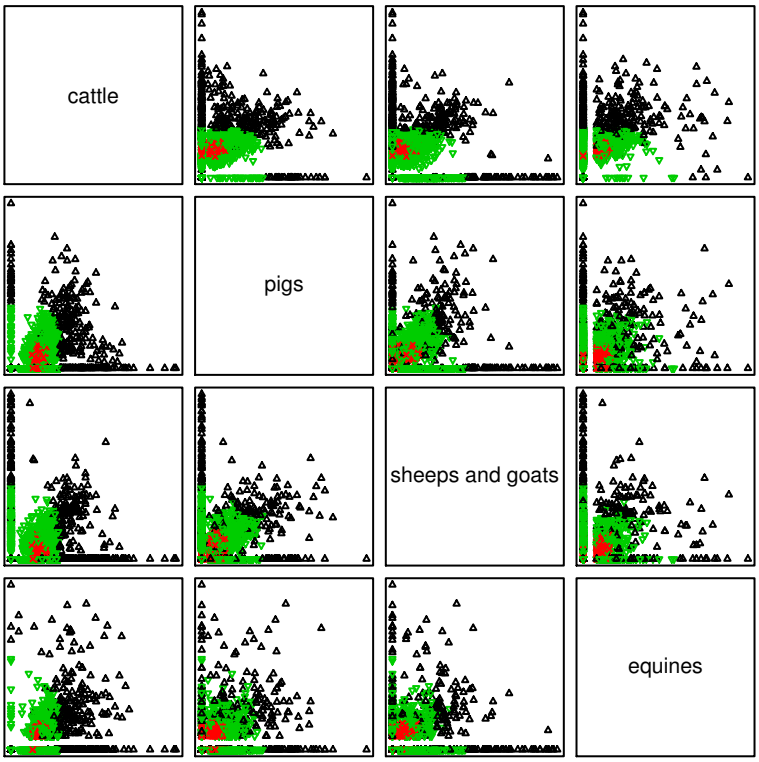

Fig. 5. Optimal partition of the population for each pair of auxiliary variables.

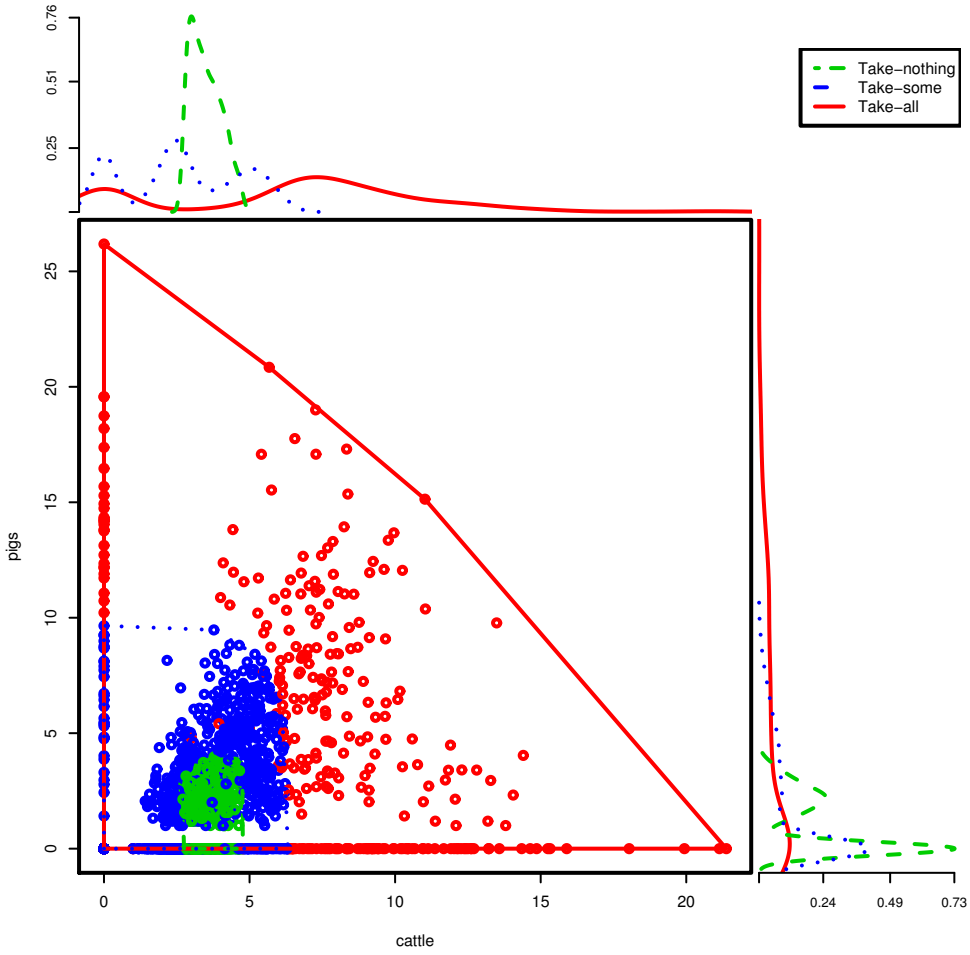

Fig. 6. Optimal partition and marginal kernel densities for cattle and pigs.

Figure 6 is an enlargement of the first subplot below the main diagonal of figure 5. This graph shows, besides the optimal partition in the three strata, the marginal kernel densities. The variability of both auxiliary variables in U_S is rather low; it is worth noting the importance of this result as this variability is the only one which affects the variance of the estimator. Roughly speaking, the variance is mostly "dumped to" the eliminated and completely enumerated strata, with the result that the variance in U_S is reduced.

The results presented here use a desired level of precision $c = 1\%$; this value has also been employed to perform the following comparisons, which show the considerable advantages of our approach in terms of sample size corresponding to the predetermined level of precision. Table 7 displays detailed results concerning some direct competitors of the cut-off design;

in particular, table 7(a) shows the sample sizes corresponding to the Hidiroglou approach, table 7(b) gives the sizes obtained stratifying the population with the K-means algorithm (Rencher 2002, sect. 14.4.1a) used as a minimizer of the variance, and table 7(c) displays the sample size corresponding to the ISTAT design introduced at the beginning of this section but setting $c = 1\%$. In the last row of table 7(a), N_C is the size of the stratum U_C obtained as the union of the four strata enumerated completely with respect to each auxiliary variable (reported in the first four rows of the table). This is one way of rendering Hidiroglou's approach, which is monopurpose and monovariate, comparable to our technique, which is multipurpose and multivariate.

Table 7(a). Sample sizes using Hidiriglou's approach

	n	N_C	N_S
$y_{1,2001}$: cattle	476.97	332	1879
$y_{2,2001}$: pigs	301.03	246	1965
$y_{3,2001}$: sheep and goats	291.29	229	1982
$y_{4,2001}$: equines	227.11	180	2031
Union	744.26	663	1548

Table 7(b). Sample sizes using the K-means algorithm

Number of strata	n
2	2193
3	2094
4	1800
5	1689
6	1259
7	1206
8	1145
9	990
10	649

Table 7(c). Sample size using the ISTAT approach

Number of strata	n
5	866

5. Conclusions

The goal of this paper consisted in proposing a framework for cut-off sampling where a model-based estimator of the unobserved part of the population plays a crucial role introducing a bias in the final estimates. The rationale for this proposal is based on the assumption that often the population distributions is highly skewed with a huge number of minor units whose weight on the population total is very small. We have discussed a formal approach for combining estimation and optimal partition of the population in three strata: census, sample and exclusion. We view this issue jointly with the multipurpose allocation of sampling units in the case where multivariate partitioning variables are available.

We have used the Simulated Annealing algorithm to minimize the number of sampling units necessary to satisfy a required precision expressed in term of MSE of the estimates of both a linear function such as the population total and a non-linear function as the ratio of a variable between two periods.

The results are encouraging: for example, for $c = 1\%$, the sample size obtained using the present approach is approximately 40 to 60% less than its direct competitors.

These outcomes also shed some light on the directions of future research in this field. In particular, we believe that attention shall be focused on the bias of the estimator with the purpose of tackling at least two issues:

- assess the robustness of the design with respect to variations of the functional form of the bias function (which here was assumed to be linear);

- use the estimation of the bias used not only for finding the optimal sample size but also for correcting the bias of the estimator used (whatever it is).

Finally, the last problem is related to the fact that the Simulated Annealing algorithm is rather slow, so that the computational burden may become unbearable when the population is large. Thus it would be the case of developing faster procedures as, for example, the Besag's (1986) Iterated Conditional Modes (ICM) algorithm.

Acknowledgements. The authors would like to thank prof. L. Biggeri (President of ISTAT) for useful comments and suggestions about the estimation of a ratio of totals and F. Piersimoni (ISTAT, Servizio Agricoltura) for providing the data.

References

Besag J. (1986) On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society B*, **48**, 259-279.

Casella, G. and Robert, C.P. (1999) *Monte Carlo Statistical Methods.* New York: Springer.

Dalenius T. (1952) The problem of optimum stratification in a special type of design, Skandinavisk Aktuarietidskrift, **35**, 61-70.

Deville J.C. and Särndal C.E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Elisson H. and Elvers E. (2001) *Cut-off sampling and estimation*, Proceedings of Statistics Canada Symposium.

Geman S. and Geman D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, **6**, 721-741.

Glasser G.J. (1962) On the complete coverage of large units in a statistical study, *Review of the International Statistical Insitute*, **30**, 28-32.

Haan J. De, Opperdoes E., Schut C.M. (1999) Item selection in the Consumer Price Index: cut-off versus probability sampling, *Survey Methodology*, **25**, 31-41.

Hansen M.H., Hurwitz W.N., Madow W.G. (1953) *Sample Survey Methods and Theory*, Vol. II. New York: Wiley.

Hidiroglou M.A. (1986) The construction of a self representing stratum of large units in survey design, *The American Statistician*, **40**, 27-31.

Horvitz D.G., Thompson D.J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663-685.

Kirkpatrick S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing, *Science*, **220**, 671-680.

Lundström S. and Särndal C.E. (1999) Calibration as a standard method for treatment of nonresponse, *Journal of Official Statistics*, **15**, 305-327.

Metropolis N., Rosenbluth A.W., Rosenbluth N.M., Teller A.H. and Teller E. (1953) Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087-1092.

Rencher A.C. (2002) *Methods of Multivariate Analysis*, 2nd edn. New York: Wiley.

Ripley B.D. (1988) *Statistical Inference for Spatial Processes.* Cambridge: Cambridge University Press.

Särndal C.E., Swensson B. and Wretman J. (1992) *Model Assisted Survey Sampling.* New York: Springer.

Sebastiani M.R. (2003) Markov random-field models for estimating local labour markets, *Applied Statistics*, **52**, 201-211 .

Stander J. and Silverman B.W. (1994) Temperature schedules for simulated annealing, *Statistics and Computing*, **4**, 21-32.

Sigman R.S. and Monsour N.J. (1995) Selecting samples from list frames of businesses. In *Business Survey Methods* (eds B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), pp. 133-152. New York: Wiley.

Statistics Canada (2001) Monthly Survey of Manufacturing (MSM), Statistical Data Documentation System, Reference Number 2101, Statistics Canada.

Winkler G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* New York: Springer.

Elenco dei papers del Dipartimento di Economia

2000.1 *A two-sector model of the effects of wage compression on unemployment and industry distribution of employment*, by Luigi Bonatti

2000.2 *From Kuwait to Kosovo: What have we learned? Reflections on globalization and peace,* by Roberto Tamborini

2000.3 *Metodo e valutazione in economia. Dall'apriorismo a Friedman ,* by Matteo Motterlini

2000.4 *Under tertiarisation and unemployment.* by Maurizio Pugno

2001.1 *Growth and Monetary Rules in a Model with Competitive Labor Markets,* by Luigi Bonatti.

2001.2 *Profit Versus Non-Profit Firms in the Service Sector: an Analysis of the Employment and Welfare Implications,* by Luigi Bonatti, Carlo Borzaga and Luigi Mittone.

2001.3 *Statistical Economic Approach to Mixed Stock-Flows Dynamic Models in Macroeconomics,* by Bernardo Maggi and Giuseppe Espa.

2001.4 *The monetary transmission mechanism in Italy: The credit channel and a missing ring,* by Riccardo Fiorentini and Roberto Tamborini.

2001.5 *Vat evasion: an experimental approach,* by Luigi Mittone

2001.6 *Decomposability and Modularity of Economic Interactions,* by Luigi Marengo, Corrado Pasquali and Marco Valente.

2001.7 *Unbalanced Growth and Women's Homework,* by Maurizio Pugno

2002.1 *The Underground Economy and the Underdevelopment Trap,* by Maria Rosaria Carillo and Maurizio Pugno.

2002.2 *Interregional Income Redistribution and Convergence in a Model with Perfect Capital Mobility and Unionized Labor Markets,* by Luigi Bonatti.

2002.3 *Firms' bankruptcy and turnover in a macroeconomy,* by Marco Bee, Giuseppe Espa and Roberto Tamborini.

2002.4 *One "monetary giant" with many "fiscal dwarfs": the efficiency of macroeconomic stabilization policies in the European Monetary Union,* by Roberto Tamborini.

2002.5 *The Boom that never was? Latin American Loans in London 1822-1825,* by Giorgio Fodor.

2002.6 *L'economia senza banditore di Axel Leijonhufvud: le 'forze oscure del tempo e dell'ignoranza' e la complessità del coordinamento,* by Elisabetta De Antoni.

2002.7 *Why is Trade between the European Union and the Transition Economies Vertical?,* by Hubert Gabrisch and Maria Luigia Segnana.

2003.1 *The service paradox and endogenous economic gorwth,* by Maurizio Pugno.

2003.2 *Mappe di probabilità di sito archeologico: un passo avanti,* di Giuseppe Espa, Roberto Benedetti, Anna De Meo e Salvatore Espa.
(*Probability maps of archaeological site location: one step beyond,* by Giuseppe Espa, Roberto Benedetti, Anna De Meo and Salvatore Espa).

2003.3 *The Long Swings in Economic Understianding,* by Axel Leijonhufvud.

2003.4 *Dinamica strutturale e occupazione nei servizi,* di Giulia Felice.

2003.5 *The Desirable Organizational Structure for Evolutionary Firms in Static Landscapes,* by Nicolás Garrido.

2003.6 *The Financial Markets and Wealth Effects on Consumption* An Experimental Analysis, by Matteo Ploner.

2003.7 *Essays on Computable Economics, Methodology and the Philosophy of Science,* by Kumaraswamy Velupillai.

2003.8 *Economics and the Complexity Vision: Chimerical Partners or Elysian Adventurers?*, by Kumaraswamy Velupillai.

2003.9 *Contratto d'area cooperativo contro il rischio sistemico di produzione in agricoltura,* di Luciano Pilati e Vasco Boatto.

2003.10 *Il contratto della docenza universitaria. Un problema multi-tasking*, di Roberto Tamborini.

2004.1 *Razionalità e motivazioni affettive: nuove idee dalla neurobiologia e psichiatria per la teoria economica?* di Maurizio Pugno.
(*Rationality and affective motivations: new ideas from neurobiology and psychiatry for economic theory?* by Maurizio Pugno.

2004.2 *The economic consequences of Mr. G. W. Bush's foreign policy. Can th US afford it?* by Roberto Tamborini

2004.3 *Fighting Poverty as a Worldwide Goal* by Rubens Ricupero

2004.4 *Commodity Prices and Debt Sustainability* by Christopher L. Gilbert and Alexandra Tabova

2004.5 *A Primer on the Tools and Concepts of Computable Economics* by K. Vela Velupillai

2004.6 *The Unreasonable Ineffectiveness of Mathematics in Economics* by Vela K. Velupillai

2004.7 *Hicksian Visions and Vignettes on (Non-Linear) Trade Cycle Theories* by Vela K. Velupillai

2004.8 *Trade, inequality and pro-poor growth: Two perspectives, one message?* By Gabriella Berloffa and Maria Luigia Segnana

2004.9 *Worker involvement in entrepreneurial nonprofit organizations. Toward a new assessment of workers? Perceived satisfaction and fairness* by Carlo Borzaga and Ermanno Tortia.

2004.10 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part I): Rational Bargaining and Justification* by Lorenzo Sacconi

2004.11 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part II): Compliance, Reputation and Reciprocity* by Lorenzo Sacconi

2004.12 *A Fuzzy Logic and Default Reasoning Model of Social Norm and Equilibrium Selection in Games under Unforeseen Contingencies* by Lorenzo Sacconi and Stefano Moretti

2004.13 *The Constitution of the Not-For-Profit Organisation: Reciprocal Conformity to Morality* by Gianluca Grimalda and Lorenzo Sacconi

2005.1 *The happiness paradox: a formal explanation from psycho-economics* by Maurizio Pugno

2005.2 *Euro Bonds: in Search of Financial Spillovers* by Stefano Schiavo

2005.3 *On Maximum Likelihood Estimation of Operational Loss Distributions* by Marco Bee

2005.4 *An enclave-led model growth: the structural problem of informality persistence in Latin America* by Mario Cimoli, Annalisa Primi and Maurizio Pugno

2005.5 *A tree-based approach to forming strata in multipurpose business surveys,* Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta.

2005.6 *Price Discovery in the Aluminium Market* by Isabel Figuerola-Ferretti and Christopher L. Gilbert.

2005.7 *How is Futures Trading Affected by the Move to a Computerized Trading System? Lessons from the LIFFE FTSE 100 Contract* by Christopher L. Gilbert and Herbert A. Rijken.

2005.8 *Can We Link Concessional Debt Service to Commodity Prices?* By Christopher L. Gilbert and Alexandra Tabova

2005.9 *On the feasibility and desirability of GDP-indexed concessional lending* by Alexandra Tabova.

2005.10 *Un modello finanziario di breve periodo per il settore statale italiano: l'analisi relativa al contesto pre-unione monetaria* by Bernardo Maggi e Giuseppe Espa.

2005.11 *Why does money matter? A structural analysis of monetary policy, credit and aggregate supply effects in Italy*, Giuliana Passamani and Roberto Tamborini.

2005.12 *Conformity and Reciprocity in the "Exclusion Game": an Experimental Investigation* by Lorenzo Sacconi and Marco Faillo.

2005.13 *The Foundations of Computable General Equilibrium Theory*, by K. Vela Velupillai.

2005.14 *The Impossibility of an Effective Theory of Policy in a Complex Economy*, by K. Vela Velupillai.

2005.15 *Morishima's Nonlinear Model of the Cycle: Simplifications and Generalizations,* by K. Vela Velupillai.

2005.16 Using and Producing *Ideas* in Computable Endogenous Growth, by K. Vela Velupillai.

2005.17 *From Planning to Mature: on the Determinants of Open Source Take Off* by Stefano Comino, Fabio M. Manenti and Maria Laura Parisi.

2005.18 *Capabilities, the self, and well-being: a research in psycho-economics,* by Maurizio Pugno.

2005.19 Fiscal and monetary policy, unfortunate events, and the SGP arithmetics. Evidence from a *growth-gap* model, by Edoardo Gaffeo, Giuliana Passamani and Roberto Tamborini

2005.20 *Semiparametric Evidence on the Long-Run Effects of Inflation on Growth,* by Andrea Vaona and Stefano Schiavo.

2006.1 *On the role of public policies supporting Free/Open Source Software. An European perspective,* by Stefano Comino, Fabio M. Manenti and Alessandro Rossi.

2006.2 *Back to Wicksell? In search of the foundations of practical monetary policy*, by Roberto Tamborini

2006.3 *The uses of the past*, by Axel Leijonhufvud

2006.4 *Worker Satisfaction and Perceived Fairness: Result of a Survey in Public, and Non-profit Organizations*, by Ermanno Tortia

2006.5 *Value Chain Analysis and Market Power in Commodity Processing with Application to the Cocoa and Coffee Sectors*, by Christopher L. Gilbert

2006.6 *Macroeconomic Fluctuations and the Firms' Rate of Growth Distribution: Evidence from UK and US Quoted Companies*, by Emiliano Santoro

2006.7 *Heterogeneity and Learning in Inflation Expectation Formation: An Empirical Assessment*, by Damjan Pfajfar and Emiliano Santoro

2006.8 Good *Law & Economics* needs suitable microeconomic models: the case against the application of standard agency models: the case against the application of standard agency models to the professions, by Lorenzo Sacconi

2006.9 *Monetary policy through the "credit-cost channel". Italy and Germany*, by Giuliana Passamani and Roberto Tamborini

2007.1 *The Asymptotic Loss Distribution in a Fat-Tailed Factor Model of Portfolio Credit Risk*, by Marco Bee

2007.2 *Sraffa?s Mathematical Economics – A Constructive Interpretation*, by Kumaraswamy Velupillai

2007.3 *Variations on the Theme of Conning in Mathematical Economics*, by Kumaraswamy Velupillai

2007.4 *Norm Compliance: the Contribution of Behavioral Economics Models*, by Marco Faillo and Lorenzo Sacconi

2007.5 *A class of spatial econometric methods in the empirical analysis of clusters of firms in the space*, by Giuseppe Arbia, Giuseppe Espa e Danny Quah.

2007.6 *Rescuing the LM (and the money market) in a modern Macro course*, by Roberto Tamborini.

2007.7 *Family, Partnerships, and Network: Reflections on the Strategies of the Salvadori Firm of Trento*, by Cinzia Lorandini.

2007.8 *I Verleger serici trentino-tirolesi nei rapporti tra Nord e Sud: un approccio prosopografico*, by Cinzia Lorandini.

2007.9 *A Framework for Cut-off Sampling in Business Survey Design,* by Marco Bee, Roberto Benedetti e Giuseppe Espa

PUBBLICAZIONE REGISTRATA PRESSO IL TRIBUNALE DI TRENTO